Published by
Tadris Matematika
IAIN Syekh Nurjati Cirebon

EduMa: Mathematics Education Learning And Teaching
December 2020, Vol 9 No 2 Page 26 – 36
https://syekhnurjati.ac.id/jurnal/index.php/eduma/index
p-ISSN: 2086-3918, e-ISSN: 2502-5209

EduMa
MATHEMATICS EDUCATION LEARNING AND TEACHING
article link: https://syekhnurjati.ac.id/jurnal/index.php/eduma/eduma/article/view/7100

Validity and Reliability Rubric of Performance Assessment Geometry Using The Many Facet Rasch Model Approach

Rivo Panji Yudha*

Faculty of Teacher Training and Education, Universitas 17 Agustus 1945 Cirebon, Indonesia

*Corresponding author: Kasepuhan, Cirebon, West Java, 45114, Indonesia. e-mail addresses: rivoyudha@yahoo.co.id

a r t i c l e i n f o

How to cite this article:

Yudha, R.P. (2020). Validity and Reliability Rubric of Performance Assessment Geometry Using The Many Facet Rasch Model Approach, 9(2), 25–34. doi:http://dx.doi.org/10.24235/eduma.v9i2.7100

Article history:
Received: 09 11, 2020

Accepted: 11 05, 2020

Published: 12, 2020

a b s t r a c t

VALIDITY AND RELIABILITY RUBRIC OF PERFORMANCE ASSESSMENT GEOMETRY USING THE MANY FACET RASCH MODEL APPROACH. The purpose of this study was to analyze the validity and reliability rubric of performance appraisal on geometry subject matter using the many facet rasch model approach through the Facets software. Data were collected from 100 small-scale students and 250 large-scale students in junior high schools using 3 raters. The performance assessment instrument in the form of a rubric is used to assess the student's process of working on the questions, each question has a different rubric. The Rasch Many Faceted Measurement Model (MFRM) is used to analyze data by looking at three aspects, namely the facet person, rater agreement, and difficulty domain using the Facet program. For the facet person, the rater separation ratio was 4.96, while the reliability of the separation index was 2.15 which indicates that the assessors are separated reliably. The stratum index is 3.21 which indicates that there are three strata of rater severity that differ statistically in the sample of these 4 raters. The rater agreement obtained the reliability of the rater separation of 0.87 and the correlation between each assessor and the other ranged between 0.40 and 0.63, indicating adequate agreement among the raters in assessing test participants with their level of competence. The Difficulty Domain on the variable maps shows that the hard to soft range is from about +1 to −1 logit.

Keywords:
Performance assessment, rubric, many-facet rasch measurement model (MFRM)

INTRODUCTION

The use of a rubric can help support a focus on education as a process. As students work shifting their learning products upwards based on the rubric scale, they learn how to improve their own learning skills simultaneously by reaching specific standards. For performance assessment, the rubric is the main tool that adds reliability, validity and transparency to the assessment. This study shows that experienced teacher-examiners are influenced by the content and nature of the rubric scale, and thus attempt to keep up.

In the assessment of performance, it is very important that the assessor knows the assessment criteria in order to provide a reliable and valid assessment (Panadero & Jonsson, 2013). Rubrics are very helpful because they provide assessment criteria in a structured format. In addition, when conducting performance assessments students are (a) reluctant to be assessed by peers who are not experts in the domain, and (b) they believe that assessment is the responsibility of the teacher (Ballantyne et al., 2002). Rubrics have the potential to alleviate both of these problems (Hafner & Hafner, 2003) and thereby increase the fairness and comfort felt with performance assessments.

The link between rubrics and learning has been explored by several researchers, with results generally indicating higher achievement and deeper learning by students who have a rubric to guide their work. Rubrics are more than just tools used to support assessors in making summative judgments. The teacher also uses the rubric as a way to provide feedback information (Nordrum et al., 2013). It is important to note that teachers are not a homogeneous group, and in the designation of modern schools it is often left to a group of para-academic specialists (Macfarlane, 2011). Students can use the rubric in a variety of ways, including self and peer assessments, and in interrogating assignment requirements (Andrade & Du, 2005; Panadero & Romero, 2014). Students use the sample rubric to: plan their responses to assignments; formative peer assessment in the classroom; and self-assessment. The teacher uses a sample rubric to provide assessment information and summative feedback.

Of course there are drawbacks too. The disadvantage most frequently cited in the literature is the fact that developing rubrics or (in the case of standard rubrics) learning to work with them is very time consuming (Diller & Phelps, 2008; Knight, 2006). In addition, class students must be trained or familiarized with the rubric before they can work with it and the process of getting used to it may take some time (Oakleaf, 2009). On the other hand, this instruction and interaction creates an awareness of the relevant criteria and a shared confidence in the competence of information, as mentioned earlier.

Several empirical studies have raised serious doubts about the validity of rubric-based performance assessments. In studies on systems thinking conducted with third and fourth year students on the undergraduate specialization in sustainability, there was a lack of valid assessments, for which many reviewers used rubrics (Habron et al., 2012). Focused holistic assessment method is used to assess student responses to each assignment. This is achieved by first developing a general scoring rubric that reflects the conceptual framework used to construct the assessment task. The general assessment rubric combines three interrelated components: conceptual and procedural knowledge of mathematics, strategic knowledge, and mathematical communication. In developing a general scoring rubric, criteria representing three interrelated components are determined for each of the four score levels (1-4).

A common problem that has existed so far is the problem of reliability for this performance assessment related to variations among assessors as a source of measurement error. Emphasizes the need to go beyond these consistency or covenant indicators and evaluate the quality of the rankings against the requirements for the invariance measure. In contrast to analyzes based on breaking down error variances into overall sources of measurement error, we highlight the importance of examining the accuracy of measurements associated with individual elements in terms such as individual rater, student, or rubric domains. In particular, the Many-Facet Rasch (MFR) model provides a useful framework in which it is possible to explore indicators of reliability and precision associated with various aspects of the assessment procedure while maintaining a focus on rater-invariant measurements.

The problem of scorer who is biased, scorer (rater) tends to be difficult to eliminate problems, personal bias. When scoring the test taker's work, there is a possibility that the scorer (rater) has a generosity error problem, meaning that the scorer tends to give high marks, despite the fact that the test taker's work results are not good. It is also possible that the scorer has a problem of severity error, meaning that the scorers tend to give low scores, even though the test taker's work results are good. Another possibility is that the scorers also tend to give moderate scores, even though in reality the test takers' work results are good and some are not. Another problem is the possibility that the scorer is interested in or sympathetic to the test taker so that it is difficult for him to give an objective score (hallo effect) (Knoch, 2009; Myford & Wolfe, 2003, 2004; Tindal, 2012; Wolfe, 2004).

The term rater variability generally refers to the variability associated with the rater's characteristics and not to the examinee's performance. In other words, rater variability is a component of undesirable variability that contributes to irrelevant construct variance in the test scores. This type of variability obscures the construct being measured and, therefore, threatens the validity and reasonableness of performance assessments (Brennan et al., 2006; Messick, 1995; Roever & McNamara, 2006; Weir, 2005), rater error (Saal et al., 1980), or rater bias (Hoyt, 2000; Johnson et al., 2009), every touch on aspects of fundamental rater variability problems.

Many mathematical concepts are represented by geometry. As stated by the National Council of Teachers Mathematics (2000) that geometric representation can help students understand the concepts of planes and fractions, histograms and scatter plots that can provide an overview of data, and coordinate graphs that relate geometry to algebra. This emphasizes the importance of geometric concepts, geometric models and spatial reasoning to interpret and describe the physical environment which can be an important tool for solving problems.

Student difficulties and teachers' needs for supplementing geometry material are the focus of study in this study. Collaborative learning can be used as a solution to improve mathematics learning, especially on the topic of geometry. This is based on the opinion of (Gonzalez & Kuenzi, 2014) who states that in teaching geometry, to activate students' knowledge in geometry learning activities is problem-based learning. The problem that is presented must be able to be absorbed and impress the students. It is the teacher's responsibility to manage interactions and learning activities that allow students to use their knowledge to solve problems.

State of the art in this research is in the design of an authentic and comprehensive performance assessment instrument using three ratings to minimize the level of subjectivity, also in terms of data analysis techniques that in previous studies still use validity and reliability calculations using classical statistical analysis, in this study the latest advances using the Manyfacet Rasch Model promises to support its use in a comprehensive manner. This study also focuses on the teacher's process of assessing work performance using instruments. Not on the results of the students working on the questions, meaning that the novelty is the process of the level of concentration between the raters in the assessment process.

METHODS

Instrument

The rubric is used as an assessment guide that describes the criteria the teacher wants in assessing or grading the results of student work. The rubric lists the desired characteristics that need to be demonstrated in a student's work accompanied by a guide for evaluating each of these characteristics. The purpose of the rubric assessment is that students are expected to clearly understand the basis for the assessment that will be used to measure student performance. Both parties (teachers and students) will have clear shared guidelines about the expected performance demands. The following is an example of a rubric for problems related to determining the properties of cubes and blocks.

Tabel 1

Rubric Problem Solving Geometry

No	Criteria	Quality Level			
		1	2	3	4
1.	Activity steps	There is no activity step	There are activity steps but they are systematic and do not lead to completion	There are systematic steps of activities but they have not yet led to completion	There is a complete systematic step about the activities that lead to completion
2	Problem Solving Process				
	Sketch	No sketch	There is a sketch but it is very imprecise and supports problem solving	There is a sketch but it does not support problem solving	There are sketches / pictures that support problem solving
	Calculation steps	There is no calculation step	The calculation steps are not systematic but the results are correct	Systematic calculation steps but wrong results	Complete systematic calculation steps and correct results

In a performance assessment, the main source of causes error in judging is the observer. The function of the observer is to make observations and give an assessment of the object being observed. This can affect the quality of performance. The attitude of subjectivity in the observation process can lead to errors in judgment so that it can reduce validity and reliability.

One of the factors that can reduce the validity of the performance assessment is biased. Bias is the teacher's mistake in interpreting student performance because it is in a group of students considered under different criteria or rated on different characteristics. If the assessment instrument that provides information is not relevant in making a decision, the instrument is invalid.

In the performance assessment assessment, a teacher must select and use fair procedures for all students regardless of cultural background, language, and gender. In addition, another factor that can cause errors in the validity of the performance assessment is the failure of the teacher to enter or provide an assessment of student performance. therefore, it takes more than one teacher to ensure that there is an agreement on values and there is no bias.

Validity and Reliability

The process of validating the observation instrument of student performance in geometry subjects is based on expert judgment through statistical measurements using CVR.

Reliability studies that involve raters are usually called inter-rater agreements or inter-rater reliability. If in the case of self-report reliability is shown by internal consistency which can be seen from one item to another that has a high correlation, then in the case of inter-rater reliability the consistency is tested for the rater. So the grain position is replaced by the person position (rater). The rater who has high agreement is seen from the position of the observed subject. If the order of subject scores from Rater A and B is almost the same, then the two raters have high agreement (Randler et al., 2011). This is because the agreement is operationalized in the form of a correlation. The rater or panelists who will be used in the assessment process are three mathematics teachers who already have certification so that later in the assessment process they can minimize the level of subjectivity.

RESULT

The results in content validation for performance assessment tools were analyzed using the content validity Lawshe where the standard of CVR validity depends on the number of Subject Matter Experts (SMEs). The CVR value must meet 0.99 so that the items can be declared valid. This applies to content validation using 3 SMEs (Lawshe, 1975, p. 568). The CVR value obtained from each item is 1 and is fully presented in the attachment. The CVI value obtained from the average CVR is 1. Based on the CVR value that exceeds 0.99, all items are declared valid (Lawshe, 1975) and are suitable for use for further research.

Next is to analyze the reliability of the instrument using the Many faceted Rasch Model approach to analyze the rating data, we obtain on a logit scale the same general interval estimate of the parameters of the aspect elements involved in the assessment (test taker performance, task difficulty level and criteria and severity level of the rater in the variable). The main benefit of the Manyfaceted rasch model (MFRM) is that, when an adequate rater match with the model is observed, rater-invariance measurements are achieved. In the context of rater-mediated mathematics performance assessment, the invariant measure implies that student achievement estimates are not

affected by which rater's score is, and the estimated rater's severity is not influenced by which problem they score. The Rasch model uses a probabilistic response distribution as a logistical function of the person and item parameters to determine unidimensional latent traits. In this study emphasis was placed on assessing rater severity. In the context of MFRM analysis, rater severity is defined as the rater's tendency to give the respondent a lower than expected average score if scores given by other raters for the same group of test takers are considered. (Myford & Wolfe, 2004).

Table 2. V Aiken
Average Score for Performance Assessment Instruments

No.	Rated aspect	Assessment criteria	Ecpert Judgment				Average CVR
			1	2	3	4	
1	Aspect conformity with indicators	1	5	5	4	4	1
		2	4	5	4	4	1
2	Writing	3	5	5	4	4	1
		4	5	5	4	5	1
		5	4	5	4	5	1
3	Language	6	5	5	5	4	1
		7	5	5	4	5	1
		8	5	4	4	4	1
4	Physical appearance	9	5	5	4	4	1
		10	5	4	5	4	1
CVI							1

In Figure 1, it can be seen that the RSR (the reliability of the separation of the assessors) is very high (0.87); is actually high enough to suggest that the observed differences in severity among raters are very reliable. In fact, the accuracy of the severity estimate is high (i.e., the standard error of the rater's severity measurement ranges between 0.03 and 0.08). The mean-square fit statistic (which ranges between 0.60 and 1.64) indicates that all raters exhibit acceptable intra-rater consistency in their assessments. The correlation of each rater with the others ranged between 0.40 and 0.63, indicating adequate agreement among raters in assessing students on their level of competence. Nonetheless, it was observed that some raters differed substantially in severity. Thus, even though their assessment shows an acceptable correlation with one of the other raters, the tendency to systematically increase or decrease the score may increase or decrease the likelihood that the test taker will pass the cut-off point.

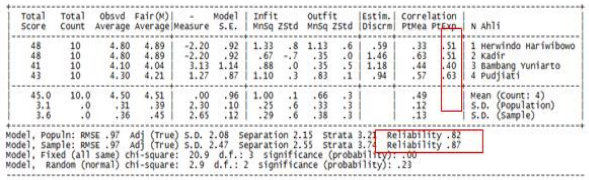

Figure 1 Results of Facets software reliability

Next is the variable map (Figure 2) showing the units of measurement (column 1) between −2 and +2 logits (log-odd-units). The rater's severity (column 2), item difficulty level (column 3), and the grading scale function (column 4) are placed on the same interval scale, creating a single frame of reference. Of particular interest to this result is column 2, which shows the variation in the severity of the rater.

Rater severity refers to an assessor's tendency to give a lower rating than other raters to take the same exam. The converse applies to assessor waivers. The variable map shows that the hard to soft range is from about +1 to −1 logit.

Obviously, the ratings aren't as bad, but the spread looks good considering there are 4 ratings. The spread range of the rater can be calculated because the separation reliability statistics are also provided by MFRM: the rater separation ratio is 4.96 while the

reliability of the separation index is 2.15 which indicates that the raters are separated reliably. The stratum index is 3.21 which indicates that there are three strata of assessor severity that differ statistically in this sample of 4 raters. In other words, as expected, regardless of the raters training and experience, they did not create a homogeneous group. The standard error of the mean of measurement is also high, namely 0.87.

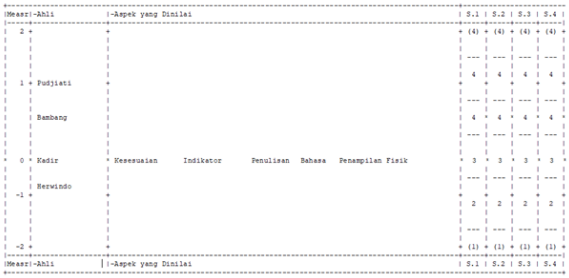

Figure 2 Expert Judgment variable map showing rater locations, indicators and thresholds

Overall, there is insufficient evidence to warrant the overall claim about the validity of the performance assessment as a class. One of the main reasons for using performance assessment is to learn things about students' knowledge and skills that cannot be learned from multiple-choice tests. However, only a few argue that there is no relationship between skills as measured by performance assessment and skills measured by multiple choice tests in the same subject. Thus, psychometrics generally look for some relationship between the two measures, but do not expect very high correlations. The ambiguity about the prediction of this relationship makes it difficult to establish simple concurrent validity arguments for a given performance assessment. As a result, performance assessments are often validated primarily on the basis of expert judgment on the extent to which the task appears to represent the construct of interest. Even here there are complications (Havarneanu, 2012). As Baxter and Glaser note, it can be difficult to design performance appraisals to measure complex understanding; as a corollary, it can be equally difficult to interpret evidence from complex performance assessments

DISSCUSSION

Current rubrics and assessment assignments, based on an analysis of student work and the use of that analysis in developing teaching strategies, provide prospective teachers with the opportunity to develop pedagogical content knowledge and skills. This is consistent with the findings of previous studies (Timmermana et al., 2011; Valli & Rennert-Ariev, 2002). Furthermore, consistent with the principles of NCTM assessment (National Council of Teachers of Mathematics, 2014), this performance-based assessment focuses on how teachers can obtain useful information about student learning so that their lessons can meet student needs.

The assessment tasks and rubrics presented in this report are aligned with local, state and national standards. These standards were at the forefront of discussion during the preparation and revision of assessment assignments and rubrics. Outside experts, particularly project evaluators and appraisal advisors, spend a lot of time reading and analyzing every word of the assignment and rubric, which has increased the validity of this grading system.

Revised assignments and scoring rubrics have proven useful for the Department of Education and Mathematics. Recognizing that teaching is complex, content-dependent, constructive, and open (Valli & Rennert-Ariev, 2002), assessment assignments and rubrics provide insight into the knowledge and skills of teacher candidate pedagogical content. The math team is satisfied that candidates who pass this assessment are ready to

start student teaching. Interviewed teacher candidates echo this sentiment and believe that, prior to teaching their students, they should be able to analyze student work and use analysis to develop lessons.

Consistent with Vogler (Vogler, 2002), assessment assignments and rubrics have resulted in some programmatic changes in teacher education programs and teaching practices in schools. The Education Office is now providing more opportunities for their students, including aspiring math teachers, to work in groups and criticize each other's work. Also, writing about mathematical concepts has been given greater emphasis in the course. Ministry of education and culture method courses now place more emphasis on demonstrating several teaching strategies. Instructors now provide more opportunities for candidates to study and analyze student work as generated by student response sample items. Candidates are encouraged to consider student responses and student prior knowledge when developing lessons and units. These changes benefit junior secondary school teacher education programs.

The Facets computer program adjusts rater performance measures for differences in rater lightness / severity, generally basing those adjustments on a single overall measure of severity for each rater. If the rater's behavior fluctuates during the rating operation, then some may question the suitability of this method to adjust ratings for differences in rater severity.

CONCLUSION

The results of the validity and reliability of the performance assessment instrument were obtained. For facet person, it was obtained that the separation ratio of the assessors was 4.96 while the reliability of the separation index was 2.15 which indicated that the assessors were separated reliably. The stratum index is 3.21 which indicates that there are three strata of assessor severity that differ statistically in this sample of 4 raters. The rater agreement obtained reliability of the rater separation of 0.87 and the correlation between each assessor and the other ranged between 0.40 and 0.63, indicating adequate agreement among the raters in assessing test participants with their level of competence. The Difficulty Domain on the variable maps shows that the hard to soft range is from about +1 to −1 logit. Obviously, the ratings aren't as bad, but the spread looks good considering there are 4 ratings.

REFERENCES

Andrade, Heidi L. and Du, Ying, "Student Perspectives on Rubric-Referenced Assessment" (2005). Educational & Counseling Psychology Faculty Scholarship. 2. Retrived from https://scholarsarchive.library.albany.edu/edpsych_fac_scholar/2

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment and Evaluation in Higher Education*. https://doi.org/10.1080/0260293022000009302

Brennan, R. L., Robert L. Brennan, & Brennan, R. L. (2006). Educational Measurement. Fourth Edition. ACE/Praeger Series on Higher Education. In *Praeger*. Retrived from https://eric.ed.gov/?id=ED493398

Diller, K. R., & Phelps, S. F. (2008). Learning outcomes, portfolios, and rubrics, oh my! Authentic assessment of an information literacy program. *Portal*. https://doi.org/10.1353/pla.2008.000

0

Gonzalez, H. B., & Kuenzi, J. J. (2014). Science, technology, engineering, and mathematics (STEM) education: A primer. In *Science, Technology, Engineering and Mathematics Education: Trends and Alignment with Workforce Needs*.

Habron, G., Goralnik, L., & Thorp, L. (2012). Embracing the learning paradigm to foster systems thinking. *International Journal of Sustainability in Higher Education*. https://doi.org/10.1108/14676371211262326

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*. https://doi.org/10.1080/0950069022000038268

Havarneanu, G. (2012). Standardized Educational Test for Diagnose the Development Level of Creative Mathematical Thinking Qualities. *International Research Journal of Social Sciences*. Retrived from http://www.isca.in/IJSS/Archive/v1/i2/5.ISCA-JSS-2012-037.php

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*. https://doi.org/10.1037/1082-989X.5.1.64

Johnson, R., Penny, J., & Gordon, B. (2009). Assessing performance: Designing, scoring, and validating performance tasks. *Journal of Educational Measurement*. 46(4), https://doi.org/10.1111/j.1745-3984.2009.00094.x

Knight, L. A. (2006). Using rubrics to assess information literacy. *Reference Services Review*. https://doi.org/10.1108/00907320610640752

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*. https://doi.org/10.1177/0265532208101008

LAWSHE, C. H. (1975). A QUANTITATIVE APPROACH TO CONTENT VALIDITY. *Personnel Psychology*. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Macfarlane, B. (2011). The Morphing of Academic Practice: Unbundling and the Rise of the Para-academic. *Higher Education Quarterly*. https://doi.org/10.1111/j.1468-2273.2010.00467.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*. https://doi.org/10.1037/0003-066X.50.9.741

Myford, Carol & Wolfe, Edward. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of applied measurement*. 4. 386-422. Retrived from https://www.researchgate.net/publication/8636147_Detecting_and_Measuring_Rater_Effects_Using_Many-Facet_Rasch_Measurement_Part_II

Myford, C. M., & Wolfe, E. W. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement*. Retrived from https://psycnet.apa.org/record/2004

-13366-007

National Council of Teachers of Mathematics. (2014). Six Principles for School Mathematics. *National Council of Teachers of Mathematics*. https://doi.org/10.1111/j.1949-8594.2001.tb17957.x

Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: Strategies for formative assessment. *Assessment and Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2012.758229

Oakleaf, M. (2009). Rubrics to assess information literacy: An examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology*. https://doi.org/10.1002/asi.21030

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. In *Educational Research Review*. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy and Practice*. https://doi.org/10.1080/0969594X.2013.877872

Randler, C., Hummel, E., Gläser-Zikuda, M., Vollmer, C., Bogner, F. X., & Mayring, P. (2011). Reliability and validation of a short scale to measure situational emotions in science education. *International Journal of Environmental and Science Education*. Retrived from https://eric.ed.gov/?id=EJ959424

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. In *International Journal of Applied Linguistics*. https://doi.org/10.1111/j.1473-4192.2006.00117.x

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.88.2.413

Timmermana, B. E. C., Strickland, D. C., Johnson, R. L., & Paynec, J. R. (2011). Development of a "universal" rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment and Evaluation in Higher Education*. https://doi.org/10.1080/02602930903540991

Tindal, G. (2012). Large-scale Assessment Programs for All Students. In *Large-scale Assessment Programs for All Students*. https://doi.org/10.4324/9781410605115

Valli, L., & Rennert-Ariev, P. (2002). New standards and assessments? Curriculum transformation in teacher education. *Journal of Curriculum Studies*. https://doi.org/10.1080/00220270110093625

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education*, 123(1), 39+.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. In *Journal of Strength and Conditioning Research*. https://doi.org/10.1519/15184.1

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46(1), 35–51. Retrived from https://psycnet.apa.org/record/2004-19990-003